A Framework of Statistical Machine Translator from English to Malayalam[1]

Mary Priya Sebastian, Sheena Kurian K and G. Santhosh Kumar
Department of Computer Science,
Cochin University of Science and Technology, Kerala, India
marypriyas@gmail.com, sheenakuriank@gmail.com, san@cusat.ac.in

Abstract: In this paper we describe the methodology and the structural design of a system that translates English into Malayalam using statistical models. A monolingual Malayalam corpus and a bilingual English/Malayalam corpus are the main resource in building this Statistical Machine Translator. Training strategy adopted has been enhanced by PoS tagging which helps to get rid of the insignificant alignments. Moreover, incorporating units like suffix separator and the stop word eliminator has proven to be effective in bringing about better training results. In the decoder, order conversion rules are applied to reduce the structural difference between the language pair. The quality of statistical outcome of the decoder is further improved by applying mending rules. Experiments conducted on a sample corpus have generated reasonably good Malayalam translations and the results are verified with F measure, BLEU and WER evaluation metrics.

Keywords: Alignment, English Malayalam Translation, PoS Tagging, Statistical Machine Translation, Suffix Separation.

1 INTRODUCTION AND RELATED WORK

Statistical Machine Translation (SMT) is one of the potential applications in the field of Natural Language Processing. SMT based on statistical method was first proposed by IBM in the early nineties [1]. Over the past years, numerous SMT systems in different foreign languages have been presented and it is observed that all of them share a common underlying structure but differ in the design of their translation model. The goal of statistical machine translation proposed here is to translate a sentence in English into the Dravidian language, Malayalam. Due to the morphological richness and intricate nature of Malayalam, as discussed in [2], very few attempts have been made to translate texts from other languages into Malayalam. To our knowledge, pure statistical machine translation from/in the Malayalam language has not been published yet.

A number of projects on machine translation from English to many Indian languages are on going in different organizations all over India [3,4]. Anglabharati, a multilingual translation system, is one among them which uses a rule-based transfer

[1] Malayalam is the native language of Kerala and it belongs to the Dravidian language family.

approach for translation. But for SMT, development of statistical methods as well as resources for training is needed. Due to the scarcity of full fledged bilingual corpus, works in this area remain almost stagnant. Experiments on statistical machine translation were carried out among many foreign languages and English. The accomplishment of an inclusive SMT system for Indian languages still remains a goal to be achieved. A work on English to Hindi statistical machine translation [5] which uses a simple and computationally inexpensive idea for incorporating morphological information into the SMT framework has been reported. Another work on English to Tamil statistical machine translation is also reported in [3]. The ideas integrated from these works have been the source of motivation and the inputs gathered from the related methodologies has facilitated in outlining the framework of the proposed SMT from English to Malayalam.

In the building process of SMT from English to Malayalam, the Malayalam corpus is subjected to some pre-processing techniques to improve the training results. To remove the insignificant alignments from the bilingual corpus a PoS Tagger is employed. Methods like suffix separation and stop word elimination from the Malayalam corpus has reduced the complexity of training. By applying order conversion rules, the English sentence is reordered to match the word order of Malayalam. Mending rules for Malayalam are designed to check the correctness of the statistical output.

The rest of this paper is organized as follows: Section 2 presents the details of the proposed architecture of the English Malayalam SMT. Some observations and the results obtained from the experiments conducted on a sample English/Malayalam corpus is discussed in Section 3. Finally, the work is concluded in Section 4.

2 PROPOSED ARCHITECTURE OF ENGLISH MALAYALAM SMT

The overall architecture of the English Malayalam SMT is given in Fig 1. In SMT, a bigram estimator [6] is employed as the language model to check the fluency of Malayalam. For the translation model, which assigns probabilities to English-Malayalam sentence pairs, IBM Model 1 training technique [7] is chosen. A variation of Beam Search method [8] is used by the decoder to work with the statistical models.

2.1 Training Phase

In the training process the sentence pairs in the bilingual corpus is converted into word aligned sentence pairs by identifying all the possible one to one mappings that exist between them. The translation parameter is estimated for the Malayalam words from these word alignments.

Setting up the corpora. Huge volumes of translated text of English and Malayalam are required to build the SMT. Malayalam corpus can be built from online Malayalam

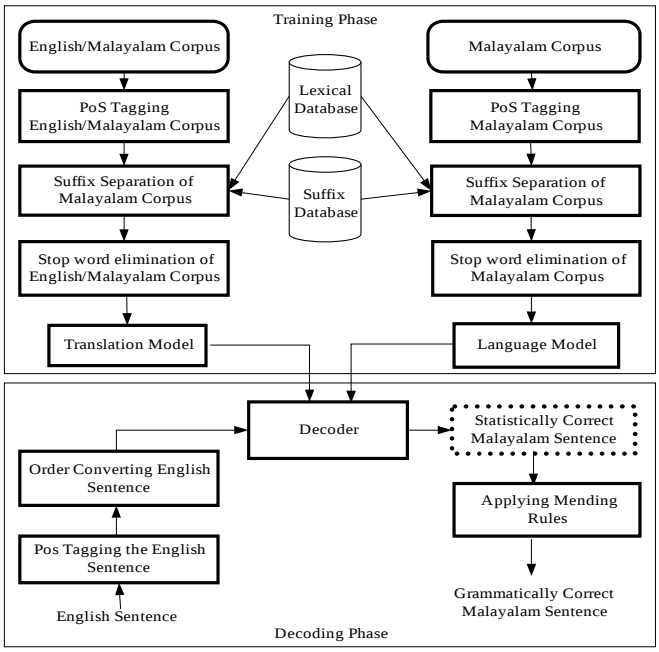

Fig 1. Overall architecture of English Malayalam SMT

newspapers and magazines. Since it is hard to find the equivalent line by line English translation, building English/Malayalam corpus is a difficult task.

PoS tagging the bilingual corpus. The most commonly used method for finding the translation probability estimate in SMT is the EM algorithm [9]. A large number of insignificant alignments have been generated when this method is adopted. The insignificant alignments carry little meaning and hence they may be eliminated from the training phase to improve the performance measure. The alignment model with PoS tagging[10] is useful in diminishing the set of alignments for each sentence pair and thereby simplifies the complexity of the training process. Here, category tags of the same type are used in tagging the words of both languages.

Suffix separation from Malayalam corpus. As discussed in [11], Malayalam language is enriched with enormous suffixes and the words appear mostly with multiple suffixes The Suffix separator is employed to extract roots from its suffixes. By incorporating a lexical database(a collection of noun roots and verb roots), a suffix database(suffixes in Malayalam) and a 'sandhi' rule generator, the functioning of the suffix separator is further enhanced, resulting in a Malayalam corpus comprising only of root words and suffixes.'ഉടെ'(ude), 'ഇൽ'(il), 'കൾ'(kal) etc are examples of suffixes separated from Malayalam corpus. Sandhi rules are framed by examining the Malayalam letter preceding the suffix in the infected form of the word. For all the Malayalam words ending in 'യുടെ'(yude), the characters before 'യുടെ'(yude) gives the root word. For example the word 'പുത്രിയുടെ'(puthriyude) is split into 'പുത്രി+ഉടെ' (puthri+ude). Certain Malayalam words, which are not in root form, still have

equivalent meaningful translations in English. The word 'അവന്റെ'(avante)' is semantically equivalent to the word 'his' in English. Even though 'അവന്റെ'(avante) has a suffix appended, it need not be suffix separated.

Stop word elimination from the bilingual corpus. Suffixes separated from the corpus are useless in the translation process. The deletion of these stop words from the corpus has brought down the complexity of the training process as well as improved the quality of the results expected from it. Similarly stop words in English language such as 'of', 'by' etc are also eliminated from the corpus before subjecting it to training.

2.2 Decoding Phase

Once the estimates for the translation parameter are obtained from training, an unseen English sentence can be translated by the decoder by applying Bayes rule [6].

Tagging the English sentence. In the decoder different syntactic tags are used to denote the syntactic category of English words. For example the sentence 'He has a car 'is tagged as He/PRP has/VBZ a/DT car/NN[2] using the POS tagger.

Order conversion. Since English and Malayalam belong to two different language families, they totally differ in their subject verb order. Order conversion rules are framed to reorder English according to the sentence structure and the word group order of Malayalam. For example, 'he ran quickly' may be translated as 'അവൻ വേഗത്തിൽ ഓടി'(avan vegathil oodi) since adverbs are always placed before verbs in Malayalam sentences.

Generating Statistically Correct Malayalam (SCM). To obtain SCM, the end product of the decoder, the order converted English sentence is split into phrases and a phrase translation table with different options of Malayalam translations is developed. Various hypotheses are created by choosing translation options and the best translation is determined by extending the hypotheses and picking the one with maximum score.

Generating Grammatically Correct Malayalam (GCM). Since SMT is trained with root words in Malayalam, the statistical outcome of the decoder lacks the required suffixes in the words generated. Hence SCM fails to convey the complete meaning depicted in a sentence. This undesirable result has been set right by applying various mending rules which helps in converting SCM into GCM. For the sentence 'I saw her', 'ഞാൻ അവൾ കണ്ടു'(njan aval kandu) is the statistical output though 'ഞാൻ അവളെ കണ്ടു'(njan avale kandu)is its correct translation. Mending Rule Applier rejoins the suffix and the word 'അവൾ'(aval) becomes 'അവളെ'(avale). For the sentence having the structure 'I/PRP saw/VBD her/PRP$', the mending rule is given as *If (PRP VBD PRP$) append the suffix 'എ' to the translation of PRP$.* Equipped with a decoder having a complete set of hand crafted rules, capable of handling all types of sentence structures, better results are obtained.

---

2 PRP, VBZ, DT and NN denote the personal pronoun, the verb in the present tense, the determiner and the noun categories respectively

3 OBSERVATIONS AND RESULTS ACHIEVED

For better training results, the corpus selected should be adequate enough to represent all the characteristics of the languages. Also, the strength and correctness of the corpus is a necessity to achieve the desired output. The sample corpus used for training includes 200 sentences with 1000 words. The experimental Malayalam corpus is built based on www.mathrubhumi.com, a news site providing local news on Kerala. The process of extending the English/Malayalam corpus is still continuing.

Table 1: Summary of evaluation results

Type of sentence	Technique	Evaluation Metric		
		WER	F measure	BLEU
Sentences in training set	Baseline + with suffix	0.3313	0. 57	0.48
	Baseline + suffix separation	0.1863	0. 78	0.69
Unseen sentences	Baseline + with suffix	0.6083	0. 26	0.22
	Baseline + suffix separation	0.4444	0.44	0.38

Evaluation metrics proposed in [12] were applied on sentences present in the training set and on totally unseen sentences. Three reference corpora were used for testing. The summary of the results are shown in Table 1. The criteria used for the evaluation are discussed below.

Word Error Rate (WER): This metric is based on the minimum edit distance between the target sentence and the sentences in the reference set.

F measure: A "maximum matching" technique where subsets of co-occurrences in the target and reference text are counted so that no token is counted twice.

BLEU: This metric is based on counting the number of n-grams matches between the target and reference sentence.

By tagging the corpus and by incorporating morphological information into the corpus, the number of alignments has reduced in the training phase. Eliminating the insignificant alignments have brought down the time and space complexity of the training process. For the annotation of the corpus with morphological information, we use an in-house parts of speech tagger for Malayalam and the Stanford POS tagger for English. The effect of suffix separation is clearly depicted in Table 1. On evaluating the results of the corpus trained without suffix separation, it was found that the final translation included many number of unwanted insertions which reduced the quality of translation. It is noted that the results of suffix separated corpus is giving better score for WER, F measure and BLEU than the one with suffixes. Even though the translations produced depicts correct meaning of the English sentence, the expected score is not met. This is due to the large number of word substitutions rather than insertions and deletions occurring in the translated sentence when compared to the reference text.

4 CONCLUSION

A system structure which can be utilized as a frame work to build a machine translation system for Malayalam using statistical models has been put forward. The method of incorporating PoS category tags into the alignment model has diminished the alignments for the sentence pairs, there by reducing the complexity of the training process. Preprocessing the corpus by incorporating suffix separation has also enhanced the quality of training. Also, application of post editing techniques like order conversion and mending rules for suffix rejoining has enhanced the outcome of the decoder. The performance of the SMT has been evaluated using WER, F measure and BLEU metrics and the results prove that the translations are of fairly good quality. This technique can be further extended and can be employed in translating any language into Malayalam by incorporating the corresponding bilingual corpus along with its order conversion rules.

REFERENCES

1. Lopez A. Statistical machine translation. *ACM Comput. Surv.,* 40, 3, Article 8, 2008.
2. A R Rajaraja Varma. *Keralapanineeyam,* Eight edition, DC books, 2006.
3. Durgesh R. Machine Translation in India: A Brief Survey. *In the Proceedings of SCALLA. Conference,* Bangalore, 2001.
4. Badodekar, S. A survey of Translation Resources,Services and Tools for Indian Languages. *In the Proceedings of the Language Engineering Conference, Hyderabad,* 2002.
5. Ananthakrishnan R, Hegde J, Bhattacharyya P, Shah R, Sasikumar M. Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. *In the Proceedings of International Joint Conference on NLP(IJCNLP08),* Hyderabad, 2008.
6. Brown P F, Pietra S A D, Pietra V J D, Mercer R L. The mathematics of statistical machine translation: Parameter estimation**.** *Comput. Linguistics, 19(2)*,pp263–31, 1993
7. Brown P F, Pietra S A D,Pietra V J D, Jelinek F, Lafferty J D,Mercer R L, Roossin P S. A Statistical Approach to Machine Translation. *Comput. Linguistics, 16(2),* pp 79–85, 1990.
8. Koehn P. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. *In Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA),* 2004.
9. Knight K. A statistical MT tutorial work book. Unpublished, http://www.cisp.jhu. edu/ws99/projects/mt/wkbk.rtf ,1999.
10. Sanchis G, S´nchez J A. Vocabulary Extension via PoS Information for SMT. *In the Proceedings of the NAACL ,*2006.
11. Sumam M I, Peter S D. A Morphological Processor for Malayalam Language. *South Asia Research, Volume27(2).* pp 173-186, 2008.
12. Stent A, Marge M, Singhai M. Evaluating evaluation methods for generation in the presence of variation. *In Proceedings of CICLing 2005, Mexico City,* pp 341-351, 2005.